

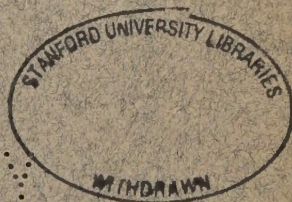
THE MEASUREMENT OF MOTOR CONTROL WITH
SPECIAL REFERENCE TO INDIVIDUAL
VARIABILITY, RELIABILITY AND
PRACTICAL IMPLICATIONS

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

BY

GAIGE BRUE PAULSEN

Stanford Library



IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
DECEMBER, 1930

3781
M66P

492183

TRAVEL AND TRAVEL

THE RELIABILITY AND CONSISTENCY OF INDIVIDUAL DIFFERENCES IN MOTOR CONTROL¹

GAIGE B. PAULSEN

Ohio University

PART 1

INTRODUCTION

THE present investigation was undertaken with a view to measuring motor control very accurately.² A large number of researches have included tests of steadiness but reports of the reliability of the test are few. Stecher has reported a reliability of .98 for a hand steadiness test when the test was divided into two sections, one comprising the morning averages of alternate days and the other the afternoon averages of the remaining days. Garfield reports a reliability of .61 on the basis of two test administrations separated by an interval of two weeks. The authors of *The Minnesota Mechanical Ability Tests* report the following reliabilities for two applications of the test separated by a three month interval: average score, .65, best score, .67, poorest score, .67 and the number of no contact holes, .64. Since Spaeth and Dunham reported a correlation of .61 between rifle shooting records and steadiness (reliability not stated), an increase in the reliability of the test from the low figures of .60 to .67 to a more acceptable value (more than .90) might increase the predictive significance of the test in that connection. Other studies employing steadiness tests consistently report low correlations with other characteristics of the individuals. One of the problems of this study was to determine

¹ This study was completed and presented at the University of Minnesota in 1930 in meeting the requirements for the Ph.D. degree. Excepting the abbreviated treatment of the literature, this article is essentially a reproduction of the thesis written at that time.

² The writer is especially indebted to Professor D. G. Paterson, who suggested the original problem and offered helpful criticism.

whether the lack of reliability of the steadiness tests is one of the main contributing factors to these low correlations.

Another problem was to determine the reason for the apparent inability of the Spearman-Brown formula to predict the reliability that resulted when the steadiness test was lengthened in the Minnesota study of mechanical ability. Their analysis showed that the reliability of the test was as great for a single hole as when the performance in all the holes was included. Such a result is not unique, since Lanier (4) found similar results for tests of musical ability and tests of elementary mechanical ability. Although he attempted no explanation of the phenomena, he made the statement that, from the standpoint of the reliability of the tests, they might just as well be half as long. If this statement be true then there is little hope of improving the reliability of our steadiness test, since repetition of measurements is one of the chief ways of improving reliability. However, the following explanation offered by the authors of the *Minnesota Mechanical Ability Tests* admits a possible basis for such increase. "Mere increase in the size of the hole may create a different psychological situation—a psychologist of the Gestalt school might term it a different configuration. If this be so, the mere addition of holes would not be comparable to increasing reliability by lengthening a test, which occurs only if added items duplicate the preceding ones."

A survey of the literature shows that the adequacy of this explanation may be rightly questioned. Holzinger and Clayton (2) have verified the predictions of the formula for evenly graded spelling material; Kelley (3) verified it for increasing the number of judges in guessing weights; Ruch, Ackerson and Jackson (6) reported verification for spelling material; Remmers, Shock, and Kelly (5) verified the formula when applied to the results of the Purdue rating scale; Lanier (4) verified it in connection with the Otis test of general intelligence. The preceding studies indicate that the Spearman-Brown formula is empirically correct not only for a wide variety of tests but also for tests which are not composed of highly homogeneous items. The present study was designed

to empirically check the adequacy of the explanation offered by the authors of the *Minnesota Mechanical Ability Tests*. If the failure of the Spearman-Brown formula was due to the lack of homogeneity of test performance, the administration of the test by using only one hole would eliminate that source of error.

APPARATUS AND METHOD

Since the *Minnesota Mechanical Ability Tests* offered data which presented a problem in measurement, a duplication of the apparatus and procedures of that study was highly desirable. Such variations as were introduced were rather insignificant and consisted in attempts to make the data more reliable and less influenced by extraneous factors. Such a procedure would have the effect of decreasing the discrepancy between the predicted and the observed reliabilities of the lack of agreement were due to the lack of controlled conditions.

The stylus, plate, and Ewald chronoscope used in the *Minnesota Mechanical Ability Tests* were used in the present study. The standard Whipple plate was mounted on a stand adjustable for height instead of being placed on the edge of a table, and the test was performed with the subject in the standing rather than the sitting position. This variation in method was adopted for two reasons. In the first place, the position is easier to control since there are not so many possibilities of securing additional bracing and support for the arm, and secondly, it was thought that the steadiness test involving the total musculature and constituting a more or less over-all measurement of the total integrated adjustment rather than the behavior of certain isolated muscle groups might have greater significance from the standpoint of vocational aptitude and general motor efficiency. A copper bottom was also added to the plate so that the contacts would be recorded if the subject inserted the stylus so far as to touch it. With this addition it was possible to observe whether the subject was receiving any support by holding the stylus on the stand upon which the plate was anchored. The complete apparatus consists of the plate, chronoscope, stylus, 3.6 volt light bulb, three dry cell batteries, and a knife

switch all wired in series. The light bulb is used to warn the experimenter when the subject is holding the stylus against the side or the bottom of the plate since in that event the recording apparatus registers only one contact. In case the subject holds the stylus against the side or the bottom, the experimenter warns against holding and the trial is continued overtime by the amount of time that the subject consumed by so doing.

The directions to the subject are demonstrated by the examiner in order to facilitate their comprehension and also to allow the subject to become somewhat at ease before the testing starts. In the first experiment, the subjects were tested with ten performances in hole eight at a single sitting and in the second experiment to be reported there were six trials at each sitting. These trials consisted of an attempt on the part of the subject to keep the stylus from touching the sides of the plate for a period of fifteen seconds after the stylus had been inserted in the hole. The trials were given at the rate of one each minute with a rest period after the fifth trial in the first study and with no rest period in the second study where only six trials were performed at a sitting.

SUBJECTS

The subjects in the first series of tests to be reported were students of laboratory psychology at the University of Minnesota. The 97 subjects of the first experiment were tested and then retested with a median time interval of one month elapsing between the tests. Every effort was made to keep the time interval constant in all cases, but some of the subjects were not retested until three months after the original test. Theoretically, the lack of absolute control of the interval would have no effect on the results if we assume that motor control is a static trait. However, the effect of this lack of control on the present results would be to magnify to some extent the effects to be reported.

The subjects for the second series of tests (results to be reported in a second article in this Journal) were 80 students of freshmen military drill at the University of Minnesota.

Each subject was tested with six trials on each of three alternate week days until ten sittings were completed. This program of testing requires continuous work on the part of the experimenter and a maximum of cooperation among the subjects. The subjects were taken from their classes (rifle marksmanship was being taught at the time) for a period of ten minutes for ten consecutive class meetings so that when the testing was completed, each of the eighty subjects had been tested with a total of sixty trials divided into ten sittings of six trials each.

RESULTS AND DISCUSSION

If the hypothesis advanced by the authors of the *Minnesota Mechanical Ability Tests* is correct then we should expect the Spearman-Brown formula to predict the added reliability that accrues when a test is lengthened by adding identical performances. Since the reliability of the steadiness test in the Minnesota study was .67 for three trials in any single hole, the reliability of the present study should be about .88 for ten trials in hole eight. However, the test-retest reliability obtained was only .73 instead of the .88 which the Spearman-Brown formula predicts for that length of test.

The first explanation of this result was that the unitary r upon which we had based our predictions was entirely in error, since this value was taken from a study having a different range of subjects and a somewhat different experimental technique. However, in order for the Spearman-Brown prediction to hold and yet have a resultant reliability of only .73 for ten trials, the unitary r would have to be about .45. Of course, the other explanation of the results would be that the Spearman-Brown formula is not adequate for this type of data.

A check of the first explanation is relatively simple since it involves the calculation of a number of unitary intercorrelations. Since there were ten trials for the first sitting, the total number of intercorrelations was 45 and they are presented in Table I.

It was soon apparent after calculation of a few of the cor-

TABLE I
Intercorrelation of Trials of Steadiness Test.
Experimental Group (N = 97)

TRIAL NO.	1	2	3	4	5	6	7	8	9	10
1		.85	.78	.78	.82	.79	.74	.71	.70	.73
2	.85		.79	.72	.80	.75	.76	.74	.74	.72
3	.78	.79		.81	.82	.72	.75	.78	.72	.73
4	.78	.72	.81		.80	.76	.72	.73	.71	.72
5	.82	.80	.82	.80		.78	.73	.81	.74	.78
6	.79	.75	.72	.76	.78		.83	.82	.81	.82
7	.74	.76	.75	.72	.73	.83		.83	.83	.82
8	.71	.74	.78	.73	.81	.82	.83		.86	.88
9	.70	.74	.72	.71	.74	.81	.83	.86		.86
10	.73	.72	.73	.72	.78	.82	.82	.88	.86	

relations that the discrepancy could not be attributed to the fact that the improper unitary coefficient had been used since all of the values in the table were over .70 and the unitary r necessary in order to make the predictions somewhat nearly correct would be .45. In fact, the average of all of the intercorrelations was .77 so that on this basis we should have expected a test-retest correlation of .97 as against the .73 that was obtained. This contrast is very suggestive of one reported in the literature, and it therefore seemed possible that the reliability could be calculated so that the predicted and empirical results would check very closely.

The reliability of .98 which was recorded in the literature resulted when two pools consisting of trials on alternate days were correlated. This procedure would tend to balance any differential rates of change in the individuals and would therefore minimize the variability which might occur in the individuals due to some factors constant for the particular test period and yet not constant for all test periods. The differential rates of change cover a period of time and it was thought possible that a study of the time element in the testing might yield results which would explain the discrepancy.

Slocombe has pointed out that the time element may be present in the intercorrelation table so that in calculating the unitary r one should be very careful to get a random sample or the predictions will not be correct. In Table I there is a very consistent effect of the time interval, *i.e.*, the trials separated by a short time interval have more in common than

those trials separated by a longer interval. It is very distinctly shown when the average of the correlation coefficients of adjacent trials is compared with the average for those trials where 1, 2, 3, 4, 5, 6, 7, and 8 trials are interpolated between those correlated. The results are presented in Table II.

TABLE II

*Effect of Time Interval on Correlation of Trials of Sitting I.
Experimental Group (N = 97)*

TRIALS INTERPOLATED	0	1	2	3	4	5	6	7	8
Av. of <i>r</i> 's	.82	.79	.78	.77	.76	.73	.73	.71	.73

In each case, there is a definite and regular decrease in the averages except where the number of interpolated trials becomes very large and the averages become unstable because of the lack of sufficient cases. In order to better visualize just what these averages represent, the reader should turn to Table I. The 0 entry of Table II is the average of all the *r*'s on the diagonal closest to the center, and the 1 entry is the average of those next to this central one and therefore has one less figure. As these diagonals recede from the center the number of interpolated trials becomes greater and the number of the *r*'s to be averaged becomes less. The effect of interpolation is so noticeable that mere inspection of the table is sufficient to establish its existence.

In view of the time effects, a check of the predictions and the empirical results was attempted in which the time relations in both were maintained fairly constant for the particular sitting. This was possible since one of the ways of determining the reliability of a test (when only one application is available) is by pooling the odd trials and even trials and correlating the two pools. In this situation the time element is reduced to a minimum since the adjacent trials are added to different pools and the transitory effects tending to influence two adjacent trials are opposed one against the other and therefore tend to cancel. If consecutive trials had not been pooled against each other, the temporary influences affecting the trials in one of the pools would produce an

added variability over that to be expected when adjacent trials are correlated.

Since the odd *versus* even technique introduces only the time element involved in adjacent trials, the unitary r on which the predictions are to be based should involve only the same time factor. The unitary r should therefore be based on the intercorrelations of adjacent trials since including all of the other r 's would make the unitary value too low. In the present intercorrelation table, the unitary r based on all the values in the table would be .77 whereas the average of the r 's of adjacent trials is .825.

Since there were ten trials in the sitting, it was possible to check the predictions up to an increase in the length of the test by five times the original (five odd trials *vs.* five even trials). The intermediate lengths were checked by computing the correlations of all possible combinations of 2 *vs.* 2, 3 *vs.* 3, and 4 *vs.* 4 trials. The summations of the trials were made in such a way that there was an absolute check on accuracy and all of the r 's were checked both in plotting and computation. The results are presented in Table III.

The predicted and empirical results agree so well that the table needs little comment. The first column gives the number of the trials cumulated and correlated. The predicted values are in the second column and the empirical results in the third column with the average values to be compared with the predicted values opposite them. The P.E.'s of the predicted values are calculated from the application of Shen's (7) formula, but they are not the limits for the average values that appear opposite the predicted values but are the limits for the individual correlations. One would expect that of all the 2 *vs.* 2 correlations fifty per cent would fall between .893 and .915; of all the 3 *vs.* 3 r 's, one half would fall between .925 and .943, and of the 4 *vs.* 4 r 's one half should fall between the limits .943 and .957. Inspection of the values reveals that a very large proportion of the r 's fall within their respective limits (10 of the 16 r 's where 8 would be expected). The empirical and the predicted results agree very closely when the time factor is held constant. The correlation of .867

TABLE III
*Empirical Results Versus Spearman-Brown Predictions.
 Odd Versus Even Trials Correlated (N = 97)*

<i>Trials Cumulated</i>	<i>Predicted Results</i>	<i>Empirical</i>
Unitary <i>r</i>		.825 ± .022
1, 3 <i>vs.</i> 2, 4		.910
2, 4 " 3, 5		.892
3, 5 " 4, 6		.866
4, 6 " 5, 7		.886
5, 7 " 6, 8		.910
6, 8 " 7, 9		.910
7, 9 " 8, 10		.899
Average 2 <i>vs.</i> 2 <i>r</i> 's	.904 ± .009	.896
1, 3, 5 <i>vs.</i> 2, 4, 6		.908
2, 4, 6 " 3, 5, 7		.919
3, 5, 7 " 4, 6, 8		.922
4, 6, 8 " 5, 7, 9		.931
5, 7, 9 " 6, 8, 10		.936
Average 3 <i>vs.</i> 3 <i>r</i> 's	.934 ± .009	.923
1, 3, 5, 7 <i>vs.</i> 2, 4, 6, 8		.948
2, 4, 6, 8 " 3, 5, 7, 9		.947
3, 5, 7, 9 " 4, 6, 8, 10		.944
Average 4 <i>vs.</i> 4 <i>r</i> 's	.950 ± .007	.946
1, 3, 5, 7, 9 <i>vs.</i> 2, 4, 6, 8, 10		.962
Value of 5 <i>vs.</i> 5 <i>r</i>	.959 ± .006	.962
1, 2, 3, 4, 5 <i>vs.</i> 6, 7, 8, 9, 10		.867

reported in the last row of the table illustrates what is to be expected in case the time factor is not controlled even within the single sitting. This correlation results when the pool of the first five consecutive trials is matched against the second five consecutive trials and is to be contrasted with the correlation resulting from the odd *versus* even correlation of the same trials.

This discrepancy is not exactly comparable to the one that might occur with learning curve data since the practice effect for steadiness is very slight within a single sitting. It indicates that there are changes in level of performance for many of the subjects, but the factors operating do not tend to produce a change which is of the same direction in the majority of the subjects. This lack of practice effect is illustrated in Table IV.

Inspection of the table reveals the fact that there is only a very slight practice effect as shown either by the means or

TABLE IV
Practice Effect in Sitting I of Steadiness Test.
Experimental Group (N = 97)

TRIAL NO.	1	2	3	4	5	6	7	8	9	10
Mean	22.47	22.53	22.35	23.07	22.14	22.14	20.78	20.43	20.14	20.16
Sigma.....	10.16	9.94	9.92	9.76	8.84	10.36	10.08	10.78	9.42	9.84

the sigmas, and that the subjects, on the average, are performing on approximately the same level throughout the ten trials. The results of this table supposedly justify our use of the Spearman-Brown formula in predicting added reliability since one of the conditions to be satisfied by the formula is that the means and the sigmas be equal, *i.e.*, there should be no practice effect. The only other condition which is imposed on the formula is that the unitary r be equal to the average of all unitary intercorrelation coefficients and since we calculated all of them we should be justified in applying the formula without reservation. With regard to the time element there has been only one caution advanced. Slocombe has pointed out that if there is such an effect one must be very careful in obtaining a unitary r based on any number of the intercorrelations less than the total number possible, since an arbitrary selection might be influenced by some constant series of changes. In spite of the fact that we have satisfied all of the restrictions imposed upon the formula, we have noted that one cannot predict the test-retest reliability on the basis of the internal consistency of the single sitting and that we cannot predict added reliability unless the time interval for the unitary r and the cumulated pools to be correlated is constant. For this reason, there must be an additional restriction imposed upon the application of the formula besides the purely statistical ones.

In case the function measured is not a relatively static one, *i.e.*, marked changes in the individual's level of performance exist, there must be a careful investigation of the time factor with a view to determining the variability to be expected because of it. The predictions must be then limited to the same time interval as was present in the establishment of the

unitary r . If such a procedure is not followed, the predicted values will overstate the reliability to be expected from the correlation of trials separated by a longer time interval than was present in the establishment of the unitary r .

RELATIONSHIP OF ODD-EVEN AND TEST-RETEST RELIABILITY

Since we found that one cannot predict the test-retest reliability on the basis of the internal consistency of the single sitting, we are faced with the problem of determining just what contribution to test-retest reliability is made by increasing the reliability of measurement within the single sitting. Lanier has stated that as far as the reliability of the music tests is concerned the tests might just as well be cut in two. If such is the case then it might be more profitable to increase the consistency by shortening the length of the individual sittings but increasing them in number. In this way we would secure a measure for each individual which would truly characterize him over a period of time. The practical significance of such a program makes it especially important that we attempt to establish some theoretical basis for the procedure.

Before any of the empirical values were established, the principles underlying attenuation suggested themselves as offering a theoretical basis for the relationships. The correction for attenuation presented a plausible means of expressing the increase in the relationship of the sittings to be expected because of some added reliability of measurement within the single sitting. However, the correction for attenuation gives the increase in the relationship to be expected when the reliability of the single sitting is perfect. If this formula is applied to the relationship between the measures secured from two sittings, then the reliability to be expected if both of the sittings measured the individual perfectly at the particular time would be the simple ratio of the test-retest reliability and the odd-even reliability. Written as a proportion this would be: Test-retest reliability : odd-even reliability (assumed to be equal in both sittings) : X (relationship with perfect odd-even r at both sittings) : 1.00 (reliability perfect at each sit-

ting). The formula in this form is very much more useful since it is possible to predict the test-retest reliability to be expected for any increase of the internal consistency in the single sitting. If the odd-even reliability for each of two sittings were .60 and the test-retest reliability were .50, and one increased the odd-even reliability to .90, a test-retest reliability of .75 could be expected, but if one made the reliability at each sitting perfect the relationship of test and retest would be only .83. This example illustrates that it may not be very profitable to increase the odd-even reliability beyond a certain point as far as the test-retest reliability is concerned since the first increase of the above sample entails a lengthening of the test by only six times the original length, whereas the relationship of .83 requires an infinite number of trials at each sitting.

The data of this research offered the possibility of making an empirical check of the above procedure since it is possible to increase the reliability of each of the sittings by increasing the number of trials upon which the measures are based and therefore to check the increases in the test-retest relationship as the number of trials are cumulated. The odd-even reliability varied from .82 to .98, depending upon whether the reliability of one or ten trials was considered so we have a fair range over which to make the test. Since the test-retest reliability for the ten trials was .73 the reliabilities to be expected for the lesser number of trials is given in Table V.

Table V was constructed with the odd-even r of .98 as a base and .73 as the empirical test-retest r . Since the increments of reliability are so very slight when the number of trials cumulated becomes fairly large, the amount added to the test-retest

TABLE V
*Test-Retest Reliabilities to be Expected on the Basis of a Varying
Odd-Even Reliability for Experimental Group (N=97)*

TRIALS USED	1	2	3	4	5	6	7	8	9	10
Odd-Even r	.83	.90	.93	.95	.96	.96	.97	.97	.98	.98
Test-Retest r	.61	.67	.69	.71	.71	.72	.72	.72	.73	.73

reliability becomes very insignificant after the test has been lengthened more than three or four times. This table would explain many of the results reported in the literature where the reliability (unspecified) is not affected by additions to the length of the test and would lead to such statements as were made by Lanier when he says that the test might just as well be cut in half as far as the reliability is concerned. The test used in this research might just as well have been cut down to three trials as far as the practical end of securing a measure which is characteristic of the individual is concerned since the other seven trials theoretically contributed (Table V) only .04 points of reliability. When the empirical result was calculated for the first three trials on the test *versus* the first three trials of retest it was found to be .67 and therefore empirically, the other seven trials contributed .06 to the test-retest reliability. This result agrees very closely with our empirical values as does the value for the one trial *versus* one trial comparison ($r = .58$ and $.58$ for two of the values as contrasted with a predicted value of .61). Other intermediate empirical values were not computed since the slight difference in the test-retest predictions would make necessary the correlation of a large number of combinations in order that the results would be able to show the fine differences. Further confirmatory evidence as to the accuracy of this procedure in predicting the increases in the test-retest reliability is offered in a thesis by Miss S. Levy. She made the predictions on the basis suggested above and the empirical results agreed very closely with the predicted values.

SUMMARY

The test-retest reliability for the steadiness test (10 tr *vs.* 10 TR) was .73, whereas the odd-even reliability for the same length of test was .98.

Test-retest reliability may not be predicted on the basis of the internal consistency of the single sitting unless the function measured is a static one.

The odd-even reliability increases according to the Spear-

man-Brown formula when the unitary r is based on adjacent trials.

The test-retest reliability *does not* increase according to the Spearman-Brown formula but the added reliability may be predicted very accurately by application of the correction for attenuation (based on odd-even reliabilities).

The means and sigmas reveal little practice effect within the trials of a single sitting, but the intercorrelation of trials shows that there are marked changes in relative position of the individuals even though these changes do not have any constant direction for the group as a whole. The changes are greater as the time interval between trials is lengthened.

A study of the increase in the reliability of the measurements by lengthening the number of sittings will be presented in an article to appear in the next issue of this publication.

BIBLIOGRAPHY

VALIDATION OF SPEARMAN-BROWN FORMULA

1. HOLZINGER, KARL J. Note on the Use of Spearman's Prophecy Formula for Reliability. *J. Ed. Psychol.*, 1923, 14, pp. 302-305.
2. HOLZINGER, K. J., AND CLAYTON. Further Experiments in the Application of Spearman's Prophecy Formula. *J. Ed. Psychol.*, 1925, 16, 289-299.
3. KELLEY, T. L. The Applicability of the Spearman-Brown Formula for the Measurement of Reliability. *J. Ed. Psychol.*, 1925, 16, 300-303.
4. LANIER, LYLE H. Prediction of the Reliability of Mental Tests and Tests of Special Abilities. *J. Exp. Psychol.*, 1927, 10, 69-113.
5. REMMERS, H. H., SHOCK, N. W., AND KELLY, E. L. An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Sscale. *J. of Ed. Psychol.*, 1927, 18, 187-195.
6. RUCH, GILES M., AKERSON, LUTON, AND JACKSON, JESSE D. An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material. *J. Ed. Psychol.*, 1926, 17.
7. SHEN, EUGENE. A Note on the Standard Error of the Spearman-Brown Formula. *J. Ed. Psychol.*, 1926, 17, 93-94.
8. SLOCOMBE, C. S. Spearman-Brown Prophecy Formula. *J. Ed. Psychol.*, 1927, 18, 125-126.
9. THURSTONE, L. L. A Note on the Spearman-Brown Formula. *J. Exp. Psychol.*, 1928, 11, 62-63.

THE RELIABILITY AND CONSISTENCY OF INDIVIDUAL DIFFERENCES IN MOTOR CONTROL. II*

GAIGE B. PAULSEN

Ohio University

SINCE it was shown in a preceding article* that steadiness records show considerable variability from time to time, we are still faced with the problem of securing an accurate index of that ability. It was apparent that certain conditions affecting the records remain relatively constant for the particular sitting but change when records are made at another time. These effects are considered as constant errors for the particular sitting but from the standpoint of repeated testings and true scores they may be considered as variable errors. If these errors behave as other variable ones, then one might secure an accurate index of steadiness by merely repeating the tests so that they have a chance to cancel each other. The reliability of our index obtained from repeated sittings should increase to the extent predicted by the Spearman-Brown formula in case the time interval is kept constant or in case the time interval has no particular effect on the relationships.

From the standpoint of securing an accurate and consistent index of the individual's steadiness, it was indicated in the previous study that one might better cut down the length of the sittings but increase them in number. The labor involved in such a testing program is much more than for the procedure where all of the trials are given at a single sitting since there is generally great difficulty in securing subjects in case repeated sittings are involved. Since the check of the Spearman-Brown formula requires that twice as many sittings be obtained as would ordinarily be required to establish a certain reliability, the testing of a sufficient number of subjects

* Part I of this study is published in the February, 1935, issue of this Journal.

(in order to secure stable correlations) at regular intervals to make a check of the formula would be almost an insurmountable task under the ordinary conditions of securing subjects. The present research was facilitated by the cooperation of the University of Minnesota Military Department, since they allowed the experimenter to test the subjects during the regular drill periods. This arrangement made it possible to test all of the 80 subjects individually during a two day period and the sittings for each subject therefore could be repeated at two day intervals. Since the subjects were tested the same hour each day and in the same order during that hour the time interval between tests was relatively constant.

The number of trials in each of the sittings was cut down to six because the previous study showed the contribution of the additional trials to the test-retest reliability to be negligible. On the basis of the earlier work this would give an odd-even reliability of .96 and it would cut down the test-retest reliability by only .01 from that possible with ten trials. It was proposed that ten sittings be given since that would make it possible to check the prediction for an increase in the length of the test by five times the original length. On the basis of the previous test-retest reliability, this increase should bring the reliability to be empirically checked to .93 and the reliability of the total number of sittings to .96.

As the testing progressed, it soon became obvious that one could not assume the various sittings to represent random samples of a more or less static ability. The practice effect for the short interval of time between the sittings was considerable and the curve of mean scores shows a consistent and steady decrease. The results are presented in Table VI.

TABLE VI
Practice Effect for Repeated Sittings. (Two Day Interval.)
Drill Group (N = 80)

SITTING NO.	1	2	3	4	5	6	7	8	9	10
Mean per Trial	18.12	14.62	13.19	13.10	12.10	10.54	9.21	8.23	7.46	9.02
Sigma	9.50	9.06	7.98	8.84	8.41	7.24	7.21	6.53	5.98	7.50
Odd-Even r	.96	.95	.96	.96	.95	.96	.95	.96	.96	.97

Both the means and the standard deviations are recorded in terms of the number of contacts per trial and not as the number of contacts for the entire six trials. In this way it is easier to compare records having a varying number of trials per sitting and the averages in this table may be compared with the results of Table IV of the preceding article. A comparison of the two shows that the average of the first sitting is below that of the previous study (18.12 *vs.* 21.62) but the standard deviations are about equal.

An inspection of Table VI shows that both the means and the sigmas decrease up to tenth sitting and then there is a regression. This regression was induced by the introduction of an emotional set in the individuals. This variation in the experimental conditions was introduced to test out the author's hypothesis that a part of the practice effect was due merely to the loss of the emotional element of the test situation. It was thought that an introduction of a reward for superior performance on the last trial might produce enough tension in the individuals to increase their records. This situation would be somewhat comparable to the experimental set-up used by Laird to determine the influence of razzing except that no distraction was introduced during the examination.

The actual testing conditions, time and method of selection of the subjects, etc., were the same except that the test was prefaced by the following remarks: "This is the last chance you will have to show your ability on this test and on the basis of today's record we are going to determine which individuals are going to be allowed to shoot on the rifle range instead of taking drill. On the basis of the first nine sittings this is your record and this is your rank in the group of 80 being tested (record sheet with name in order of rank with score opposite.) Everybody will have a chance to win since the five men that show the greatest improvement, *i.e.*, change the most ranks, will be able to shoot as well as the five men that make the best score on the basis of today's records." These remarks were followed by a few statements showing

what a good chance each man had to win and with the final caution that he should do his best. Throughout the last sitting there were no further remarks made and the experimenter maintained a passive and neutral attitude.

The effects of this procedure on the means and the sigmas has already been pointed out, but it will be noted that there is no appreciable effect on the odd-even reliability. In fact, the odd-even reliability is not even affected by the changes in the means and the sigmas throughout the series of trials. This indicates that the individuals change (for the slight time interval involved between the trials) relatively the same amount whether it is in the first sitting or the last one, and that the odd-even reliability does not reflect the changes in the experimental conditions because of the way in which it is obtained (two trials under very nearly the same conditions pooled against each other).

The change in experimental conditions and the differential changes in the individuals are reflected in the intercorrelation of sittings which is presented in Table VII.

TABLE VII
*Intercorrelation of Sittings of Steadiness Test. Six Trials per Sitting,
Two Day Intervals (N=80)*

SITTING NO.	1	2	3	4	5	6	7	8	9	10
1		.80	.65	.65	.64	.61	.67	.59	.53	.63
2	.80		.88	.75	.71	.62	.72	.64	.66	.69
3	.65	.88		.76	.73	.68	.76	.64	.61	.70
4	.65	.75	.76		.85	.72	.77	.72	.72	.77
5	.64	.71	.73	.85		.81	.81	.78	.79	.87
6	.61	.62	.68	.72	.81		.85	.83	.73	.76
7	.67	.72	.76	.77	.81	.85		.79	.73	.83
8	.59	.64	.64	.72	.78	.83	.79		.87	.78
9	.53	.66	.61	.72	.79	.73	.73	.87		.77
10	.63	.69	.70	.77	.87	.76	.83	.78	.77	

Inspection of this table shows that those sittings that are separated by the greater amount of time are less related than those which are consecutive. The exact amounts of decrease

in the average of the correlation for the sittings separated by a varying number of sittings is shown in Table VIII.

TABLE VIII
Effect of Time Interval and Practice on the Correlation of Sittings
(Drill Group, $N = 80$)

Sittings Interpolated	0	1	2	3	4	5	6	7	8
Av. of r 's	.82	.75	.73	.72	.71	.67	.65	.61	.62

This table is very similar to the one for trials (Table II of previous study) and the same consistent decrease is evident although the effect must be attributed to two causes (time interval and practice) whereas practice did not affect the intercorrelation for trials.

It is interesting to note the effect of the motivation introduced at the tenth sitting. The experimenter thought that the addition of competition on the last sitting would tend to re-introduce some of the emotional element into the situation and therefore make the last sitting have more in common with the first few sittings than the next to the last one. Table VI presented evidence to show that the average performance was considerably disturbed by the element of competition and the intercorrelation table shows the increased relationship of the tenth sitting with the first sitting over that of sitting 9 *vs.* sitting 1. The same relation holds when sitting two and three are included in the comparison with nine and ten and one is probably justified in concluding that the dropping out of the emotional element is a considerable factor in both the practice effect and in the differential changes in the performance of the individuals which is reflected in the intercorrelation table.

SPEARMAN-BROWN PREDICTIONS FOR REPEATED SITTINGS

The presence of systematic changes in the scores of the individuals as the sittings progress make it improbable that the Spearman-Brown formula may be applied without taking into

account the effect of the time interval. If this is the case, then we will not be able to state the added consistency that will accrue because of the increased number of sittings unless we confine the predictions to the time interval upon which the unitary r is based, and unless the unitary r is based upon a uniform time interval instead of on the average of all of the intercorrelations. When we make our predictions in that manner the following table results:

TABLE IX

Empirical Results versus Spearman-Brown Predictions for Odd versus Even Sittings (N=80)

SITTINGS CUMULATED	PREDICTED RESULTS	EMPIRICAL RESULTS
Unitary r	.821 $\pm$.024	.821 $\pm$.024
1, 3 vs. 2, 4		.873
2, 4 " 3, 5		.921
3, 5 " 4, 6		.904
4, 6 " 5, 7		.924
5, 7 " 6, 8		.866
6, 8 " 7, 9		.917
7, 9 " 8, 10		.922
Average 2 vs. 2 r 's	.902 $\pm$.015	.904
1, 3, 5 vs. 2, 4, 6		.942
2, 4, 6 " 3, 5, 7		.953
3, 5, 7 " 4, 6, 8		.924
4, 6, 8 " 5, 7, 9		.941
5, 7, 9 " 6, 8, 10		.944
Average 3 vs. 3 r 's	.932 $\pm$.011	.941
1, 3, 5, 7 vs. 2, 4, 6, 8		.945
2, 4, 6, 8 " 3, 5, 7, 9		.957
3, 5, 7, 9 " 4, 6, 8, 10		.948
Average 4 vs. 4 r 's	.948 $\pm$.009	.949
1, 3, 5, 7, 9 vs. 2, 4, 6, 8, 10		.961
Value of 5 vs. 5 r	.958 $\pm$.007	.961
1, 2, 3, 4, 5 vs. 6, 7, 8, 9, 10		.852

Inspection of this table shows that the predictions and the empirical results agree very well when the time interval is maintained constant. The P.E.'s are the limits for the indi-

vidual correlations and not for the average values which they are opposite. In making the detailed comparison of the individual correlations with the predicted values, we note that of the 2 *vs.* 2 *r*'s only two of the seven are within the limits of the P.E. of the predicted value even though the average value is only .002 from the predicted one. The remainder of the individual *r*'s agree so well with the predicted values that, of the sixteen possible comparisons, nine are within the limits of the P.E.'s of the predicted *r*'s.

The fact that differential changes are involved in the levels of performance for the different individuals as the sittings progress is again shown by the fact that the correlation of the first five and the last five sittings is only .85. On the basis of the reliability of five sittings as against one, the reliability predicted for that cumulation was .96. One might have expected that reliability had we not shown previously that the performance on the successive sittings shows a considerable practice effect when the repetitions of the test are spaced by one day intervals only. With the introduction of a longer time interval between the successive sittings, the element of adaptation to the test and testing situation might have been eliminated so that the successive tests of the individual's ability might have been considered as random samples. If such conditions were fulfilled and steadiness could be considered as having no constant direction of change for the individuals tested, the predicted reliability on the basis of two successive records would be an accurate statement of the reliability to be expected for any length of test considered. The preceding analysis of the data secured from various measurements of steadiness have shown such ideal conditions of measurement to be quite imperfectly fulfilled, since it seems that steadiness can not be assumed to be constant nor can the testings be spaced to eliminate the practice effect. Under such conditions it is important that the predictions by means of the Spearman-Brown formula be limited to the time interval involved in the determining of the unitary *r*.

SUMMARY OF EXPERIMENTAL DATA

There is a very marked practice effect on both the means and the standard deviations of the repeated sittings. One of the factors contributing to this practice effect is the adaptation to the test situation since an introduction of an emotional set in the individuals on the last trial disturbs their performance markedly.

The same discrepancy between the test-retest reliability and the odd-even reliability is evident for all of the sittings. The internal consistency of the single sittings does not change with the reduction in the means and the sigmas but is relatively constant for all of the sittings ($r = .96$). This value may be contrasted with the test-retest reliability of about .82 which is the average value for the correlation of consecutive sittings.

The reliability of the sittings increases according to the Spearman-Brown formula in case they are cumulated by the odd-even method.

The reliability of the sittings *does not* increase according to the Spearman-Brown formula in case the time interval for the cumulations and for the unitary r is not constant.

The presence of systematic changes in the repeated sittings makes the procedure outlined for securing a very consistent index of steadiness very laborious since reliability does not increase according to the Spearman-Brown formula.

Steadiness cannot be considered a static function and no single measure or group of measures occupying a short space of time may be considered to represent more than a fair index of the ability of the subjects.

THE MEASUREMENT OF DYNAMIC FUNCTIONS—AN INDEX OF TRAIT VARIABILITY

Most of the measures of psychological traits imply that the function measured is static. The errors of measurement due to such factors as fatigue, shifts of attention, emotional set or attitude, physiological and other factors are considered to be perfectly random in their operation or non-operative. However, these factors are of such nature that they tend to

persist during a certain period of measurement and since their effects may be absent in measures secured at another time, they may be considered as constant errors for a given sitting. The presence of such constant errors at a particular sitting would make it impossible to secure a measure of a trait at one particular time which would necessarily characterize the individual over a longer period. Any trait which behaves in this fashion may be termed dynamic and if there were no changes of constant direction in the trait as repeated samples were made, high reliability could be obtained by a series of measurements secured at different times. This procedure would convert constant errors present only at particular times into variable errors over a period of time.

The above mentioned factors may not be operative for some measurements and considerable work in mental measurement has demonstrated the relative constancy of the IQ. Of course, lack of perfect consistency is to be expected for any psychological function, but the errors of measurement of intelligence are largely attributed to errors of the measuring instrument such as lack of homogeneity and length of test. The fact that differences in IQ's at different testings show only a slight tendency to increase as the length of time between the testings is increased has been the evidence which supports the majority opinion with regard to the constancy of the IQ, *i.e.*, errors in the measuring instrument account for practically all of the observed fluctuations.

In measuring a static trait under ideal conditions, it would be possible to secure a perfectly reliable index of the individual's ability by merely increasing the length of the test at a single sitting or by splitting up the necessary number of trials into several or many sittings. There would be no difference in the results achieved by the two procedures if the individual's ability were assumed to be constant. It is perfectly obvious, however, that if an individual's ability should be perfectly measured at one time and subsequently some factor conditioning it should be temporarily or permanently altered, our theory of measurement would no longer be ade-

quate. Perfect consistency of the function measured is implied in practically all psychological measurements and many text books of statistics have a statement to the effect that there is theoretically no difference as far as added reliability is concerned whether the test is doubled in length or whether two different applications of the standard length test are made. Psychologists have not taken the word "theoretically" into account and have proceeded to apply statistical techniques to data where the constancy of the trait has not been proven.

The fact that intelligence has been found to be approximately static and independent of changes operating over a relatively long period of time has led to the assumption that the same statistical techniques may be applied to other psychological functions without an examination of the constancy of the particular trait. However, the preceding data have shown what is to be expected when such assumptions are uncritically made in measuring motor control or steadiness. Predictions as to increased reliability as a result of lengthening the test were found to be greatly in error and correction of observed relationships for attenuation on the basis of the reliabilities reported would lead to an underestimation of the true correlations between the abilities measured. This underestimation is due to the fact that reliabilities reported were too high, in which case the correction for attenuation does not increase the existing relationships by a sufficient amount. It is apparent that the degree to which a trait is static or dynamic is very important for our manipulation of test results and for our interpretation of the relationships discovered.

No single quantitative statement exists for the purpose of characterizing a measured function with reference to its degree of variability or consistency. The consistency of the particular function under consideration is dependent not only upon the variability inherent in the trait but also upon the degree to which the errors of observation have been eliminated (Lengthening the test, etc.). If we could correct all test-retest reliability coefficients on the basis of the odd-even r , then they would be reduced to a common basis of comparison. A

method for such correction which relates odd-even and test-retest coefficients was outlined in a previous article. There we demonstrated that this correction was empirically justified for test material which is as homogeneous as repeated test trials. It will be remembered that test-retest reliability was corrected for attenuation on the basis of the internal consistency of the single sitting and that the correction for perfect internal consistency merely reduces to the ratio of the test-retest r and the odd-even r , *i.e.*, the maximum relationship to be expected from a test *vs.* retest r for a particular trait = $\frac{\text{test-retest } r}{\text{odd-even } r}$. Thus if the odd-even reliability of a test were .50 and the test-retest r were .40, the maximum relationship to be expected between any two sittings for the trait would be .80, $\left(\frac{.40}{.50}\right)$. One would be justified in increasing the odd-even reliability in the above sample, but if the odd-even r had been .95 instead of .50 the maximum relationship to be expected would be only .42. This is an increase in the test-retest r of only .02 and represents the fruits of an increase in the test length up to an infinite number of trials per sitting. If both the test-retest and the odd-even r were equal then the correction would bring the relationship up to 1.00 and it would be theoretically possible to secure a perfectly consistent index of the trait by merely extending the length of the test at a single sitting.

The more practical use of the formula is in connection with increases of the internal consistency where the odd-even r is not increased to 1.00 but to some intermediate value. In that case the correction is given by the proportion: test-retest r : odd-even r :: X (corrected test-retest r): proposed odd-even r . If the odd-even r of the previous example were increased from .50 to .80 then we might expect an increase in the test-retest r to .64. This increase in the odd-even reliability coefficient could be attained by an increase of the test by four times its present length. If the test having an odd-even reliability of .95 and a test-retest reliability of only .40 were increased by

four times the original length, the test-retest reliability would only be increased to .41 and certainly that increase (.01) in the test-retest r would not justify the additional labor involved in such a testing program. Under these conditions it would be advisable to increase the number of sittings, since the test-retest reliability coefficient should increase according to the Spearman-Brown formula, and a repetition of the test at four different sittings should increase the reliability to .73. The preceding examples have demonstrated that the calculation of these "corrected" test-retest coefficients furnishes definite information which may serve as a basis of a program for securing very reliable measures of the individuals tested.

It is proposed that this "corrected" test-retest coefficient be termed Coefficient of Trait Variability. This new index indicates the degree to which dynamic changes in individuals reduce the correlation from the perfect correlation (1.00) which might be expected if errors of observation were entirely eliminated. It is the observed test-retest reliability coefficient corrected for the errors of measurement and represents the correlation to be expected if the measures of the two sittings correlated are perfect. As was stated previously, the identity of the odd-even and the test-retest correlations would reveal that the trait measured is a static one since the correlation of the results of perfect measurement at each of two sittings would be 1.00. However, any discrepancy in the odd-even and the test-retest correlations indicates the presence of variability in the individuals tested and the Coefficient of Trait Variability states the maximum correlation to be expected from two different sittings even if each of the sittings has perfect internal consistency.

SUMMARY AND CONCLUSIONS

The present investigation was undertaken to ascertain the reasons why the steadiness test, which had been used in the Minnesota Mechanical Ability Research Project, did not increase in reliability (test-retest) as the number of trials (in different holes) was cumulated. The main result is the dis-

covery that the test was not at fault but that the individual's ability changed from time to time as a result of factors constant for a particular sitting but variable from sitting to sitting. In other words, trait variability and not errors of measurement are involved. Therefore, increase in the length of the steadiness test *at a particular sitting* merely serves to emphasize errors constant for that sitting instead of eliminating them. Hence the Spearman-Brown formula based on odd-even reliability coefficients is not applicable with reference to test-retest reliability. As a result the correlation between two different sittings will not increase as predicted.

Coefficients of Trait Variability if computed for various traits would probably show a great range and one would probably find steadiness to be quite variable in comparison with intelligence, since the latter has been shown to be relatively constant. The present research study has shown steadiness to be so variable that the maximum relationship which one could expect to find between two sittings, each of infinite length, would be about $+.80$. On the other hand, the test-retest reliability coefficients for measurements of intelligence are actually much higher and hence the Coefficient of Trait Variability for that function would probably be somewhere in the range from $.90$ to $.98$. Similarly, the determination of Coefficients of Trait Variability for the entire range of psychological (physiological and growth functions also) functions would add materially to our present meagre knowledge regarding trait variability in the individual.

In recognizing that many psychological traits are dynamic (those for which we cannot secure a characteristic index at one sitting), we may seem to have played into the hands of those critics of experimental psychology who insist that mental processes in individuals are so changeable that they cannot be accurately measured. We have admitted the variability of traits within the individual but we have not been content to rest with that statement. We have offered a quantitative index of trait variability and on the basis of that index have suggested a means for increasing the reliability of the mea-

surement of a dynamic function. Essentially, the suggested procedure emphasizes the importance of increasing the number of sittings rather than perfecting measurements within a single sitting.

If the constant errors present at each sitting behave in a random manner when the sittings are repeated, the reliability of the measurement of a dynamic function may be increased by merely increasing the number of sittings. The reliability for a series of sittings should theoretically increase according to the predictions made by the Spearman-Brown formula and, therefore, the added reliability secured by this method should increase rapidly. Unfortunately, the complete verification of the Spearman-Brown formula for this procedure with the steadiness test could not be expected because the sittings were secured at such short intervals that the combined effect of practice and adaptation was considerable. However, the procedure was theoretically justified since the empirical reliability coefficients for odd-even sittings satisfied the Spearman-Brown predictions. In this manner we have shown that many tests which might be abandoned as hopeless, judged from the standpoint of test-retest reliability alone, can be justified if administered several different times.

The most important implication of this thesis for psychometric theory and practice is the increased emphasis which should be placed upon repeated examinations of individuals at different times rather than the completion of testing procedures within a given examination period. The thesis should also serve to clear up much of the obscurity and misunderstanding which has resulted from observed discrepancies between odd-even and test-retest reliability coefficients. Finally, there is the demonstration that the Spearman-Brown Prophecy Formula, although mathematically correct, is applicable only to data derived from the measurement of static traits.

VITA

AUTHOR—Gaige Brue Paulsen.

BORN—Hollandale, Wisconsin, July 22, 1905.

ACADEMIC HISTORY—Graduated from South High School, Minneapolis, 1923. B.A., *magna cum laude*, 1927, University of Minnesota. Entered the Graduate School, University of Minnesota, September, 1927; received the Ph.D. in December, 1930, major Psychology, minor Medicine (Neurology). Teaching Assistant in Psychology, University of Minnesota, 1927-29.

PROFESSIONAL EXPERIENCE—Instructor in Psychology, University of Minnesota, 1929-30.

PRESENT POSITION—Assistant Professor, Department of Psychology, Ohio State University, Athens, Ohio, 1930-.

